Proposal

The objective of this experiment is to optimize the performance of two text classification models using Stochastic Gradient Descent. In addition, this experiment will propose fine tuning the SGD based classification models using the Grid-Search approach. To conduct this, various NLP preprocessing tasks and transformations will be performed on Global Terrorism dataset obtained from Kaggle, originally housed and maintained by University of Maryland was utilized. The aim is to determine whether Grid Search can be used to find parameters that optimize the SGD classifier, and determine that the SGD learning algorithm can serve as an optimizer for text classification models. The main package that will be used is sklearn, almost an all-encompassing library that allows us to perform various methods including but not limited to logistic regression and SVM. It also lets us perform gridsearch and provides us with tools to score and evaluate models built.

Evaluation

- Accuracy: To calculate the percentage of samples in the test set that the classifier correctly labeled, measured by calculating TP+TN/TP+FP+FN+TN
- Precision: Indicates how many incidents were relevant, and measured by TP/ (TP+FP)
- Recall: Indicates how many of the relevant items were identified, and measured by TP/ (TP+FN)
- F1 combines the precision and recall to give a single score. It is the harmonic mean of the precision and recall and calculated by (2 × Precision × Recall)/ (Precision Recall)

Data source

https://www.kaggle.com/datasets/START-UMD/gtd?select=globalterrorismdb_0718dist.csv

Feature: Summary of terrorism attack

Target: Attack type

GitHub

https://github.com/Chim515

Schedule

11.28	Proposal
12.1	Preprocessing
12.5	Modeling
12.11	Presentation + Write Up Report
12.12	Tweak and improvements